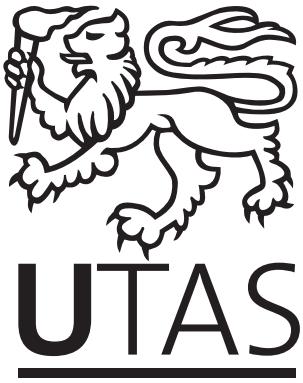
ENTANGLEMENT, INVARIANTS, AND PHYLOGENETICS

by

Jeremy G Sumner, B.Sc. Hons (Tas)

Submitted in fulfilment of the requirements
for the Degree of Doctor of Philosophy

School of Mathematics and Physics
University of Tasmania
December, 2006



I declare that this thesis contains no material which has been accepted for a degree or diploma by the University or any other institution, except by way of background information and duly acknowledged in the thesis, and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due acknowledgement is made in the text of the thesis.

Signed: _____
Jeremy G Sumner

Date: _____

Signed: _____
Jeremy G Sumner

Date: _____

The following people contributed to the publication of work undertaken as part of this thesis.

Entanglement invariants and phylogenetic branching [59].
Jeremy G Sumner (75%), Peter D Jarvis (25%).

Using the tangle: a consistent construction of phylogenetic distance matrices [60].
Jeremy G Sumner (80%), Peter D Jarvis (20%).

We the undersigned agree with the above stated proportion of work undertaken for each of the above published (or submitted) peer-reviewed manuscripts contributing to this thesis.

Signed: _____

Peter D Jarvis
Supervisor
School of Mathematics and Physics
University of Tasmania

Date: _____

Signed: _____

Larry Forbes
Head of School
School of Mathematics and Physics
University of Tasmania

Date: _____

ABSTRACT

This thesis develops and expands upon known techniques of mathematical physics relevant to the analysis of the popular Markov model of phylogenetic trees required in biology to reconstruct the evolutionary relationships of taxonomic units from biomolecular sequence data.
The techniques of mathematical physics are plethora and have been developed for some time. The Markov model of phylogenetics and its analysis is a relatively new technique where most progress to date has been achieved by using discrete mathematics. This thesis takes a group theoretical approach to the problem by beginning with a remarkable mathematical parallel to the process of scattering in particle physics. This is shown to equate to branching events in the evolutionary history of molecular units. The major technical result of this thesis is the derivation of existence proofs and computational techniques for calculating polynomial group invariant functions on a multi-linear space where the group action is that relevant to a Markovian time evolution. The practical results of this thesis are an extended analysis of the use of invariant functions in distance based methods and the presentation of a new reconstruction technique for quartet trees which is consistent with the most general Markov model of sequence evolution.

ACKNOWLEDGEMENTS

First and foremost my thanks go to my supervisor Peter Jarvis. Not only for having the insight to take on this novel work and his outstanding knowledge of mathematical physics, but also for being a true friend and good bloke.

These people have all played their own special role in bringing this thesis to fruition: Michael Sumner, Robert Delbourgo, Patrick McLean; William Joyce and the Physics department of the University of Cantebury; Mike Steel and the organisers of the New Zealand phylogenetics meeting; Rex Lau, Lars Jermiin, Michael Charleston and SUBIT; Alexei Drummond; Simon Wotherspoon (for giving me such a hard time), Malgorzata O'Reilly, Jim Bashford, Giuseppe Cimo, Stuart Morgan, Isamu Imahori and Graham Legg; Mum, Dad and Kate; Keith, Tim, Sarah, Wazza and Beans.

A special mention for my high school maths teacher Mr. Rush, who used to laugh when I continually interrupted his classes with: "That's all very well, Mr. Rush, but how is this going to help me lay bricks?"

Here, as it draws to its last Halt, if anywhere, might both Gentlemen take joy of a brief Holiday from Reason. Yet, "Too busy," Mason insists, and "Far too cheerful for thah'," supposes Dixon.

Mason and Dixon
Thomas Pynchon

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

The rationale of this thesis is taken from a remarkable analogy between the stochastic models used to infer phylogenetic relationships in mathematical biology and the structure of multiparticle quantum physics. There is a direct relationship between Feynman diagrams that describe the interactions of sub-atomic particles and phylogenetic trees that graphically represent the evolutionary relationship between taxonomic units. A Feynman diagram gives the graphical representation of creation and annihilation events of particle interactions. A taxonomic unit may be any biomolecular unit such as a gene, an amino acid or base pair, and the time evolution of these molecular units is modelled stochastically under a Markov assumption. Techniques which reconstruct the evolutionary history of molecular units from present observations are based on these models. Given the correct framework, these Markov models and the formalism of multiparticle quantum mechanics can be put into a mathematical correspondence. This is a very useful observation because phylogenetics is a relatively new mathematical problem (for example see the classic paper by Felsenstein [19]) whereas the mathematics of particle physics has been studied for over a century. (For an outstanding introduction to the history of theoretical particle physics see [47], and for a comprehensive introduction to mathematical physics see [61].) Given that there is a mathematical connection between the two problems it would certainly be unfortunate to see results that have been obtained in physics re-derived independently in the context of phylogenetics. This thesis looks at a particular aspect of quantum systems known as *entanglement* and shows that measures of entanglement can be utilized to improve the reconstruction of phylogenetic relationships.

We will need to be clear that the probabilities associated with quantum systems and those of phylogenetic models arise in quite a different scientific way. Quantum mechanics is a probabilistic theory because the theoretical predictions give the correct statistical behaviour regarding the outcomes of particular experiments. The theoretical predictions can be used to infer (incredibly accurately) the distribution of results for many repetitions of the same experiment. (For a popular discussion of the amazing accuracy of quantum theory see Feynman's discussion of the magnetic moment on the electron as predicted from quantum electrodynamics [22].) Since quantum theory is (and should be) seen

as a *theory* of nature there has been argument for many decades on how to interpret this probabilistic aspect of quantum theory. This argument raises quite profound scientific and philosophical issues which, thankfully, we will not be concerned with in this thesis. Models of phylogenetics are exactly that – *models*, and should not be seen as being theories of nature. No one would argue that the time evolution of molecular units follow the Markov model of phylogenetics in detail, but rather that these models are the best (tractable) approximation that give us recourse to establishing properties of phylogenetic history. Primarily the points of interest are the branching structure of the evolutionary history and also the evolutionary distance (or time) between branching events.
After we have made the mathematical analogy between quantum theory and the Markov model of phylogenetics, we will concentrate on only a small part of what can be done using techniques known in mathematical physics. We will focus on the study of entanglement invariants and their generalization to the phylogenetic case [59, 60]. There is potential for concentrating on other techniques such as Lie algebra symmetries [6] and the analysis of the path integral formulation [31, 32], but these techniques will not be explored here.
The distance based technique has been used in phylogenetics as a tree building algorithm following the discovery that it is possible to calculate a distance from the observed sequences that is consistent with the Markov model. This distance function is a well defined mathematical object known as a group invariant function and is used in quantum physics to quantify and test for the phenomenon of entanglement. Entanglement is a general property that can exist in many different physical systems and the invariant function used as a distance measure in phylogenetics is used to quantify entanglement for only the most elementary case. Hence, it seems astute to investigate what the next most complicated types of entanglement correspond to in phylogenetics.

Theoretical outcomes of the thesis

We present a group representation theoretic analysis of the Markov model of phylogenetic trees. Specifically this formalism is used to construct all the one-dimensional representations of the (appropriately defined) Markov semigroup. These one-dimensional representations occur as polynomials in the (discrete) probability distributions predicted from the Markov model which we coin *Markov invariants*. We establish the connection between these one-dimensional representations and that of *phylogenetic invariants* [11, 15, 20, 55] and pairwise distance measures [25, 40]. This representation theoretical approach touches upon existing techniques and can be incorporated into known algorithms to give novel results and insights to the problem of phylogenetic reconstruction. The main theoretical outcome of the thesis is this use of representation theory. We will also develop the theory of invariants of the general linear group on a tensor product space and show how to infer existence of these invariants in different cases. We develop a procedure for computing the explicit form of these invariant functions, firstly developed for the general linear group and then generalized to the Markov semigroup.

Practical outcomes of the thesis

We study a group invariant function, well known in quantum physics as the *tangle*, in the context of phylogenetics. The tangle is used in physics to give a measure of the amount of entanglement between three qubits. Qubits are two state objects in quantum physics and correspond in phylogenetics to a probability distribution on two states. In phylogenetics the classic example is to use the DNA as a state space and hence the case of four state objects is of interest. To this end we have generalized the tangle to the case of three and four character states. This is a new result that to the best of the author's knowledge was previously unknown. Having successfully generalized the tangle we investigate how the tangle can be used to construct improved phylogenetic distance matrices. Additionally we study a set of Markov invariants which exist for the case of phylogenetic quartet tree. In the case of the evolution of four taxa there are three possible historical evolutionary relationships. We show that these Markov invariants can be used to distinguish these three cases under the assumption of the most general Markov model. It is expected that the use of the tangle to construct distance matrices and using the Markov invariants to distinguish the three possible quartets will lead to improvements of the reconstruction of phylogenetic relationships from observed biomolecular data.

Structure of the thesis

Chapter 2 begins by introducing the mathematical material needed to understand the results presented in this thesis. This includes a short introduction to group representation theory, group characters and tensor product; a presentation of the Schur/Weyl duality and the Schur functions; a definition of group invariant functions and their relation to one-dimensional representations. The chapter ends with several relevant examples of invariants of the general linear group.
Chapter 3 begins with a light speed introduction to the formalism of quantum mechanics, the concept of entanglement and mathematical analysis thereof using group invariant functions. The Markov model of phylogenetic trees is then developed in its usual presentation, followed by a change of formalism which makes apparent the analogy between phylogenetic trees and multiparticle quantum systems. The chapter ends with a detailed analysis of the mathematical analysis of the invariant functions when evaluated upon a phylogenetic tree.
Chapter 4 gives a review of phylogenetic distance measures and shows how the tangle invariant function used to analyse three qubit entanglement can be generalized to the phylogenetic case and used to improve popular distance measures. This is done by defining the branch lengths of a phylogenetic tree, reviewing the standard measure known as the log det and then using the tangle invariant to give a consistent distance measure for the case of quartets.
Chapter 5 returns to the mathematical detail of Chapter 2 and derives in-

variant functions that are more closely relevant to the Markov model of a phylogenetic tree. This is done by first defining the Markov semigroup. The invariant functions of the general linear group are rederived using a technique which is generalized to derive the Markov invariants. Finally we examine the structure of the Markov invariants on a phylogenetic tree. In particular we concentrate on the quartet case where there exists four Markov invariants which can be used to distinguish between the three possible quartet trees.